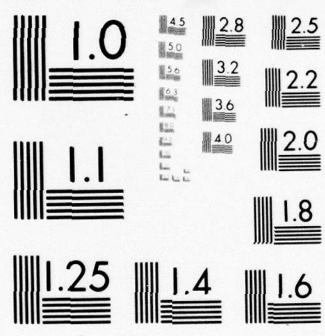
AD-A069 815 MINNESOTA UNIV MINNEAPOLIS DEPT OF PSYCHOLOGY F/G 5/10
RELATIONSHIPS AMONG ACHIEVEMENT LEVEL ESTIMATES FROM THREE ITEM--ETC(U)
APR 79 G G KINGSBURY, D J WEISS N00014-76-C-0627
UNCLASSIFIED RR-79-3 NL

1 OF 1
AD
A069815

END
DATE
FILMED
7-79
DDC



MICROCOPY RESOLUTION TEST CHART

LEVEL

RELATIONSHIPS AMONG ACHIEVEMENT LEVEL ESTIMATES FROM THREE ITEM CHARACTERISTIC CURVE SCORING METHODS

AD A069815

G. Gage Kingsbury
and
David J. Weiss

RESEARCH REPORT 79-3
APRIL 1979

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

This research was supported by funds from the Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, Office of Naval Research, Army Research Institute, and Air Force Human Resources Laboratory, and monitored by the Office of Naval Research.

79 06 12 165

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(9)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 79-3
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): (6) Relationships Among Achievement Level Estimates from Three Item Characteristic Curve Scoring Methods.
5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) G. Gage Kingsbury and David J. Weiss
8. CONTRACT OR GRANT NUMBER(s): (15) N00014-76-C-0627
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Minnesota, Minneapolis, MN 55455
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (16) P.E.: 61153N PROJ.: RR042-04; (17) T.A.: RR042-04-01; W.U.: NR150-389
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, VA 22217
12. REPORT DATE: (11) April 1979
13. NUMBER OF PAGES: 28
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (14) RR-79-3 (12) 37p.
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
18. SUPPLEMENTARY NOTES: This research was supported by funds from the Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, Army Research Institute, Air Force Human Resources Laboratory, and Office of Naval Research, and monitored by the Office of Naval Research.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

latent trait test theory	achievement testing	testing
item response theory	computerized testing	tailored testing
response-contingent testing	adaptive testing	programmed testing
individualized testing	sequential testing	automated testing
	branched testing	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study compared achievement level estimates from three item characteristic curve (ICC) scoring methods using the one-, two-, and three-parameter ICC models. The three scoring methods were maximum-likelihood normal, maximum-likelihood logistic, and Owen's (1975) Bayesian scoring method. Data included all possible response patterns from a hypothetical five-item test, as well as response patterns from live administration of a conventional and an adaptive achievement test. For the conventional and adaptive test data, correlations among achievement level estimates were examined as a function of test length.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified 406 024
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LB

parameter value

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Results for all data sets showed a high degree of similarity among θ estimates for the one- and two-parameter data, with slight decreases in correlations as information on the discrimination parameter was used in scoring. When the third ("guessing") parameter was used in scoring the item response data, correlations among θ estimates were reduced, particularly for the adaptive test data. The data also showed an increasing tendency for the maximum-likelihood methods to result in convergence failures as the third parameter of the ICC was used in scoring. In general, however, the adaptive test data were less likely to result in convergence failures than were the conventional test data. The data also illustrated how each of the three scoring methods tend to utilize ICC parameter information in arriving at θ estimates and the relationships of these estimates to a number-correct scoring philosophy. Advantages and disadvantages of each of the scoring methods are discussed. It is suggested that future research examine the relative validities of scoring methods and model combinations.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS

Acknowledgements

Adaptive and conventional test data utilized in this report were obtained from volunteer students in General Biology, Biology 1-011, at the University of Minnesota during fall quarter 1976 and winter quarter 1977; appreciation is extended to these students for their participation in this research. The cooperation of Kathy Swart and Norman Kerr of the General Biology staff in providing access to the students is also deeply appreciated.

Technical Editor: Barbara Leslie Camm

Relationships Among Achievement Level Estimates from Three Item Characteristic Curve Scoring Methods

With the advent of computerized instruction and testing, and the concurrent reduction in costs of minicomputer systems, it has become feasible to use item characteristic curve (ICC) response models to estimate students' achievement levels, based on responses to classroom tests. This feasibility has been demonstrated recently in an experimental context (Bejar, Weiss, & Kingsbury, 1977; Reckase, 1977), and computer programs for implementing these scoring methods have been made available (Bejar & Weiss, 1979). These technological advances should be paced by theoretical advances if perspective is to be maintained and the maximum possible return from advancing technology is to be insured.

When ICC response models are employed within a classroom situation, estimates of the achievement level of any student may be obtained in a number of different ways (Bejar & Weiss, 1979). The two most widely used scoring methods are the maximum-likelihood (M-L) estimators (Lord & Novick, 1968) and the Bayesian estimators (Lindgren, 1976). Estimates obtained by a M-L procedure will be asymptotically consistent and unbiased. The property of consistency implies that as the number of items answered by an individual increases toward infinity, the difference between the M-L estimate of a student's achievement level ($\hat{\theta}$) and the actual value of the parameter (θ) will approach zero. Therefore, as a test becomes very long, an estimate of the achievement level will approach the actual achievement level. The property of unbiasedness implies that if several M-L estimates of an achievement level are made, the mean of the estimates will equal the actual θ value. These properties are highly desirable from a statistical point of view.

Although estimates obtained using a Bayesian procedure (e.g., Owen, 1975) allow the incorporation of prior information into the achievement level estimation process, they are somewhat biased. This bias in the Bayesian achievement level estimates has been demonstrated by McBride and Weiss (1976) in a series of computer simulations. In each case the Bayesian scoring method was shown to provide θ estimates with average values different from the true θ levels that gave rise to the response pattern. Thus, individuals with a high true achievement level received an ability estimate that was lower than the true θ value, and individuals with a low true θ level received a θ estimate somewhat higher than the true value. The bias increased as the estimated θ became more discrepant from the true θ level.

Both M-L and Bayesian scoring methods allow the use of all of the information contained in the testee's responses to all the items in the test in order to arrive at the final estimate of the testee's achievement level. However, the Bayesian algorithm devised by Owen (1975) is somewhat affected by the order of the items in the test; that is, scoring the responses in a different order will result in a different estimate of trait level (Sympson,

1977). On the other hand, the M-L estimators are independent of the item order. In general, in a test of finite length, a single response pattern may receive differing achievement level estimates solely as a function of the differences between the scoring methods.

Samejima (1969) has noted that M-L estimates for individuals will differ as a function of the underlying response model. More importantly, though, she has pointed out that ordering of individuals' trait level estimates will change as a function of the response model assumed in the scoring method. Bejar and Weiss (1979) have also noted, within a two-parameter ICC model, that a difference in the ICC scoring method used will result in different trait level estimates for the same pattern of responses to the same test items. These investigators used all possible response patterns in a hypothetical five-item test to illustrate differences among three different methods for estimating trait levels; however, there is some question whether the differences found within the hypothetical data set used will generalize to live-testing data sets. According to ICC response theory, not all response vectors are equally likely. Because the hypothetical data sets used in the Samejima (1969) study and the Bejar and Weiss (1979) study were highly improbable--each possible response pattern occurred once--results from real data sets may reflect different levels of similarity among the results of different ICC scoring methods.

If differences in ordering of individuals as a function of the ICC scoring method are found in real data sets, such results will have direct consequences for educators who are preparing to implement a testing system utilizing ICC theory and procedures. In an educational situation, the ordering of individuals according to their responses on tests is of paramount importance. For this reason, it is important to determine the degree of disparity in achievement level estimates based on the different methods of scoring item responses using ICC theory. Similarly, since test response patterns can be scored by using one, two, or three of the parameters describing the ICC, different levels of similarity among θ estimates may be obtained by different scoring methods using each of the models.

The recent experimental applications of adaptive testing strategies in educational settings (e.g., Bejar, Weiss, & Gialluca, 1977; Brown & Weiss, 1977) may open the way to the use of shorter, more precise individualized tests in future classrooms. Since the Bejar and Weiss (1979) and Samejima (1969) data suggest that short tests may result in differences among achievement levels estimated by different scoring methods, it is imperative that the implementation of adaptive testing systems be accompanied by a knowledge of the differences among the achievement level estimates resulting from different scoring strategies for adaptively administered achievement tests. A beginning toward the development of this knowledge is simply the recognition that differences do exist among the various scoring methods and that these differences may have an impact on rankings of the individual students in the classroom. The present study was designed to investigate these differences through additional analyses of the data reported by Bejar and Weiss (1979) and Samejima (1969) and through analysis of data from the administration of conventional and adaptive tests.

Method

The three scoring methods described by Bejar and Weiss (1979) were compared across three different ICC response models. The three scoring methods were (1) maximum likelihood using a normal probability function (M-L normal), (2) maximum likelihood using a logistic probability function (M-L logistic), and (3) Owen's Bayesian scoring method using a constant prior with a mean of 0 and a standard deviation of 1.0. The three ICC response models were (1) the one-parameter model, in which test items differ only in terms of their difficulties (Rasch, 1960); (2) the two-parameter model, in which items may differ in terms of their difficulties and discriminations (Lord & Novick, 1968); and (3) the three-parameter model (Lord & Novick, 1968), in which items may differ in terms of difficulties, discriminations, and "guessing" parameters.

Test Data

Data used were from three different sources: (1) the hypothetical test and the structured set of response patterns used by Bejar and Weiss (1979), (2) a conventional classroom achievement test, and (3) a computer-administered adaptive achievement test.

Hypothetical response patterns. Using the example provided by Bejar and Weiss (1979), achievement level estimates were obtained for each possible response pattern to a hypothetical five-item test for which the parameters for each of the three response models were assumed to be known. The parameter values for the hypothetical test using the three-parameter model are shown in Table 1. All 32 possible response patterns were generated for these five items (see Table 2). Since M-L scoring methods cannot score response patterns with all items answered correctly or all items answered incorrectly, analyses were confined to the 30 response patterns scorable by all three scoring methods.

Table 1
Item Parameters for a Hypothetical Five-Item Test
Assuming a Three-Parameter ICC Response Model

Item	Discrimination (a)	Difficulty (b)	Lower Asymptote (c)
1	1.00	-2.00	.10
2	1.50	-1.00	.10
3	1.00	0.00	.10
4	1.50	1.00	.10
5	1.00	2.00	.10

Conventional test. Data were obtained from the administration of a conventional classroom achievement test to a group of 200 undergraduate college students in an introductory biology course at the University of Minnesota. Estimates of the parameters of the three-parameter ICC model were available for 39 of the 55 items administered in this particular examination (see Bejar, Weiss, & Kingsbury, 1977).

The item parameter estimates were obtained using a method operationalized by Urry (1976). The procedure performs a direct conversion of the classical item parameters to obtain estimates of the discrimination (a) and difficulty (b) parameters and uses the value that minimizes a χ^2 statistic as an estimate of the "guessing" (c) parameter. Estimates are further refined by an ancillary correction procedure. Estimates of the parameter values for this examination were based on the responses of approximately 1200 people to each item. Final parameter estimates are shown in Appendix Table A.

Adaptive test. To determine whether the process of adapting a test to an individual's level of achievement might also affect the extent to which the different scoring methods yielded similar achievement level estimates for a group of individuals, additional data were obtained from the live administration of a computerized stratified adaptive (stradaptive) test. Utilizing the item pool from which the conventional test was drawn, this test was administered to a group of 200 volunteer students from the same biology course (Bejar, Weiss, & Gialluca, 1977).

The parameter estimates for the items in the stradaptive item pool were obtained from previous administrations of conventional classroom examinations. The ICC item parameter estimation procedure was the same as that used for the conventional test. The number of individuals on which the parameter estimates were based ranged from 638 to 998, depending on the original time of administration of the item. The parameters of the items in the stradaptive item pool are shown in Appendix Table B. The stradaptive test used a variable termination rule which terminated the test when an individual's ceiling stratum (Weiss, 1974, p. 46) had been identified. Test lengths actually taken by individuals varied from a minimum of 9 items to the maximum of 50 items.

Scoring and Analysis

Hypothetical test. Each of the 32 response patterns was scored by each of the three scoring methods (M-L normal, M-L logistic, and Bayesian) using the parameter values from Table 1. This represented an application of the three-parameter model. In order to use the two-parameter model, each of the response vectors was again scored with each scoring method; but the value of c for each item was set to zero (values of a and b for each item remained the same as in Table 1). To apply the one-parameter model, each response pattern was again scored by each scoring method; but the value of a for each item was set equal to 1.00, and the value of c was set to zero (values of b again remained as in Table 1).

To determine the extent to which the scoring method employed in achievement level estimation affected the rank ordering of the 32 response patterns, two analyses were performed. First, for each response model, differences among the scoring methods were examined by determining for each pair of scoring methods (1) the number of response patterns which were given different rankings, (2) the magnitude of the greatest difference in ranking, and (3) the average difference in ranking across all response patterns. Secondly, the degree of agreement among the scoring methods was quantified by obtaining values of Kendall's Tau (a rank order correlation coefficient) between achievement level estimates obtained from each pair of scoring methods within each response model. To the extent to which these correlations differed from

1.0, the scoring methods involved may be said to give divergent rankings of the same response patterns.

Conventional and adaptive tests. Conventional and adaptive test response patterns from the 200 subjects were scored by each of the three scoring methods at various points in the test. Scores were obtained after each three-item block in the test. Thus, this procedure produced scores based on the administration of 3 through 39 items in the conventional test and 3 through 48 items in the adaptive test, in increments of 3 items. This scoring was done first under the assumption of the three-parameter model, using the available item parameter estimates from Appendix Tables A and B. To investigate scoring by the two-parameter model, the scoring procedure described above was again employed (i.e., all response patterns were scored by each of the three scoring methods at each of a number of different test lengths). However, the parameters were edited so that although a and b for each item remained the same as in Appendix Tables A or B, c for each item was set to zero. Scoring by the one-parameter model was also done at 3-item increments for each test; but item parameter values were edited so that a for each item was set equal to 1.00, c for each item was set equal to zero, and b for each item remained as in Tables A or B.

Correlations were then calculated separately for the one-, two-, and three-parameter data between achievement level estimates generated by each pair of scoring methods at each of the 13 different test lengths between 3 and 39 items for the conventional test, and at each of the 16 different test lengths from 3 to 48 items for the adaptive test. To the extent that any correlation differed from 1.0, it might be said that at that particular test length the two scoring methods gave achievement level estimates that differed by more than a linear transformation.

Results

Hypothetical Test

One-parameter model. The achievement level estimates obtained for each of the possible response patterns from each of the scoring methods, assuming a one-parameter ICC response model, are shown in Table 2. The response patterns in which all items were answered correctly [1,1,1,1,1] and in which all items were answered incorrectly [0,0,0,0,0] have been omitted because the M-L estimates for these response patterns are positive and negative infinity, respectively. To make the comparison among scoring methods easier, the estimates have been ordered in terms of the ranking of the Bayesian achievement level estimates.

For the one-parameter model, the Bayesian achievement level estimates differed from the M-L normal estimates in rank order for 17 of the 30 response patterns. The average difference in ranking of a response pattern between the two methods was .43. The greatest difference in ranking between scores derived from the two models was a difference of 1.5 ranks.

The Bayesian estimates differed from the M-L logistic estimates in rank order for 28 of the 30 response patterns. The average difference in rank order was 2.07. The largest difference in ranking was 4.5 positions. This result

was confounded, however, by the large number of tied ranks obtained by the M-L logistic scoring method; there were only 4 unique scores for the 30 response patterns. By contrast, the Bayesian method gave unique θ estimates to all 30 response patterns.

Table 2
Achievement Level Estimates and Rank Orders for
Bayesian and Maximum-Likelihood (M-L) Scoring Methods
Assuming a One-Parameter ICC Response Model

Response	Bayesian		M-L Normal		M-L Logistic	
Pattern[a]	Estimate	Rank	Estimate	Rank	Estimate	Rank
1,1,1,0,1	1.05	1	1.59	2	1.61	3
1,1,0,1,1	1.00	2	1.38	3	1.61	3
1,1,1,1,0	.97	3	1.62	1	1.61	3
1,0,1,1,1	.84	4	1.09	4	1.61	3
0,1,1,1,1	.58	5	.79	5	1.61	3
1,1,0,0,1	.54	6	.69	6	.51	10.5[b]
1,0,0,1,1	.47	7	.51	8.5[b]	.51	10.5
1,0,1,0,1	.42	8	.51	8.5	.51	10.5
1,1,0,1,0	.38	9	.51	8.5	.51	10.5
1,1,1,0,0	.34	10	.51	8.5	.51	10.5
0,1,0,1,1	.27	11	.29	12	.51	10.5
1,0,1,1,0	.27	12	.33	11	.51	10.5
0,0,1,1,1	.24	13	.20	14	.51	10.5
0,1,1,0,1	.21	14	.26	13	.51	10.5
0,1,1,1,0	.07	15	.07	15	.51	10.5
1,0,0,0,1	.01	16	-.07	16	-.51	20.5
0,0,0,1,1	-.03	17	-.20	17	-.51	20.5
0,1,0,0,1	-.15	18	-.26	18	-.51	20.5
0,0,1,0,1	-.16	19	-.29	19	-.51	20.5
1,0,0,1,0	-.21	20	-.33	20	-.51	20.5
1,0,1,0,0	-.33	21	-.51	22.5	-.51	20.5
1,1,0,0,0	-.33	22	-.51	22.5	-.51	20.5
0,0,1,1,0	-.34	23	-.51	22.5	-.51	20.5
0,1,0,1,0	-.35	24	-.51	22.5	-.51	20.5
0,1,1,0,0	-.47	25	-.69	25	-.51	20.5
0,0,0,0,1	-.53	26	-.79	26	-1.61	28
0,0,0,1,0	-.78	27	-1.09	27	-1.61	28
0,0,1,0,0	-.97	28	-1.38	28	-1.61	28
1,0,0,0,0	-1.02	29	-1.62	30	-1.61	28
0,1,0,0,0	-1.05	30	-1.59	29	-1.61	28

[a]The response patterns [0,0,0,0,0] and [1,1,1,1,1] are not included because M-L estimates cannot be obtained for these response patterns.

[b]Ties were assigned the average of the ranks that the tied estimates would span if they were not tied.

The ranks of the M-L normal estimates differed from those of the M-L logistic method for 28 of the 30 response patterns. The average difference

in rank order was 2.00, and the maximum difference in ranking was 4.5. Again, the small number of unique ranks assigned by the M-L logistic method partially accounted for this difference; the M-L normal method gave unique θ estimates to 24 of the 30 response patterns.

It is evident from these data that using the one-parameter model, the three scoring methods resulted in different θ estimates. Although there were only relatively small differences in the rank ordering of the θ estimates between the Bayesian and the M-L normal methods, all θ estimates generated by the Bayesian method were uniformly closer to zero than those of the M-L normal method. The differences were particularly large at the extremes, where the differences were as much as .50 score units on the achievement metric for the [1,1,1,0,1] and [0,1,0,0,0] response patterns. The tendency of the Bayesian θ estimates to be closer to zero was also evident in comparison to the M-L logistic method. However, because of the tendency of the M-L logistic method not to provide different θ estimates for different response patterns, differences approaching .50 units were evident between the two methods for response patterns obtaining θ estimates near the mean (e.g., response pattern [1,0,0,0,1]).

Using the one-parameter model, the M-L logistic scoring method resulted in different θ estimates for different numbers of items answered correctly. Thus, θ estimates of 1.61 were obtained for all response patterns in which only 4 items were answered correctly; θ estimates of .51 were given to all response patterns in which 3 items were answered correctly; θ estimates of -.51 were obtained for all patterns with 2 correct answers; and θ estimates of -1.61 were assigned to all patterns with only 1 correct answer. It should be noted that the items were all of differing difficulties (see Table 1). Thus, the one-parameter M-L logistic scoring method provides θ estimates based on the number of items answered correctly, but does not take into account the difficulties of the items; all response patterns with the same number-correct score will result in the same θ estimates, regardless of whether easy or difficult items are answered correctly. This property of the one-parameter M-L logistic scoring method is the basis for the use of number-correct score in the Rasch (1960) one-parameter logistic ICC model. By contrast, both the M-L normal and Bayesian scoring methods resulted in different θ estimates for items of differing difficulty; in these scoring methods the difficulties of items answered correctly or incorrectly are taken into account in estimating θ levels.

Two-parameter model. The estimates of achievement level for all the possible response patterns (except [0,0,0,0,0] and [1,1,1,1,1]) for the two-parameter response model are shown in Table 3; for these data the Bayesian estimates differed from the M-L normal estimates in terms of rank order in 16 of 30 instances. The average difference in rank position between the two methods was .65; the maximum difference in the ranking of the two methods was a difference of 3 positions.

The Bayesian estimates differed from the M-L logistic estimates in rank order for 28 of the 30 response patterns, and the average difference in rank position was 1.93. The maximum difference in rank was 4.5 positions.

The M-L normal estimates differed from the M-L logistic estimates in terms of rank order for 28 of the 30 response patterns, and the average

difference in rank position was 1.63. The largest discrepancy in the rankings was a difference of 4.5 positions.

Table 3
Achievement Level Estimates and Rank Orders for Bayesian and Maximum-Likelihood (M-L) Scoring Methods Assuming a Two-Parameter ICC Response Model

Response	Bayesian		M-L Normal		M-L Logistic	
Pattern[a]	Estimate	Rank	Estimate	Rank	Estimate	Rank
1,1,0,1,1	1.09	1	1.42	2	1.60	2
1,1,1,1,0	1.08	2	1.63	1	1.60	2
1,1,1,0,1	.93	3	1.24	3	1.19	4.5
0,1,1,1,1	.64	4	.93	4	1.60	2
1,1,0,1,0	.63	5	.78	5	.84	7
1,0,1,1,1	.62	6	.61	6	1.19	4.5
1,1,0,0,1	.51	7	.60	7	.46	11.5
0,1,0,1,1	.41	8	.50	8	.84	7
1,0,0,1,1	.39	9	.30	11	.46	11.5
1,1,1,0,0	.31	10	.42	9	.46	11.5
0,0,1,1,1	.30	11	.13	14	.46	11.5
0,1,1,1,0	.28	12	.39	10	.84	7
1,0,1,1,0	.23	13	.17	13	.46	11.5
0,1,1,0,1	.17	14	.23	12	.46	11.5
1,0,1,0,1	.11	15.5[b]	.03	15	.00	15.5
0,0,0,1,1	.11	15.5	-.13	17	-.46	19.5
0,1,0,1,0	.00	17	-.03	16	.00	15.5
1,0,0,1,0	-.06	18	-.17	18	-.46	19.5
0,0,1,1,0	-.11	19	-.30	20	-.46	19.5
0,1,0,0,1	-.15	20	-.23	19	-.46	19.5
1,0,0,0,1	-.24	21	-.39	21	-.84	24
0,0,1,0,1	-.28	22	-.50	23	-.84	24
1,1,0,0,0	-.29	23	-.42	22	-.46	19.5
0,0,0,1,0	-.38	24	-.61	25	-1.19	26.5
0,1,1,0,0	-.42	25	-.60	24	-.46	19.5
1,0,1,0,0	-.58	26	-.78	26	-.84	24
0,0,0,0,1	-.64	27	-.93	27	-1.60	29
0,1,0,0,0	-.89	28	-1.24	28	-1.19	26.5
0,0,1,0,0	-1.06	29	-1.42	29	-1.60	29
1,0,0,0,0	-1.16	30	-1.63	30	-1.60	29

[a]The response patterns [0,0,0,0,0] and [1,1,1,1,1] are not included because M-L estimates cannot be obtained for these response patterns.

[b]Ties were assigned the average of the ranks that the tied estimates would span if they were not tied.

As in the case of the one-parameter model, it was again apparent that the three scoring methods resulted in different estimates of achievement levels. Estimates obtained from the Bayesian method showed the same tendency toward more moderate estimates (i.e., estimates closer to zero) that was exhibited

using the one-parameter model. This result occurred when the Bayesian scoring method was compared with either of the M-L scoring methods. The magnitude of the discrepancies between the Bayesian estimates and the M-L normal estimates was almost exactly the same as with the one-parameter model. Comparison between the Bayesian estimates and the M-L logistic estimates was again made difficult by the fact that the M-L logistic method sorted the 30 response patterns into only 9 different achievement levels. However, differences between the estimates appeared to be greater for response patterns which received extreme achievement estimates than for those which received moderate estimates.

The observation that the M-L logistic method yielded 9 different achievement levels indicates that the number of correct responses is no longer a sufficient description of the M-L logistic achievement level estimate using the two-parameter model. In fact, as the data in Table 3 indicate, the sufficient indicant of the M-L logistic achievement level estimate using the two-parameter model was the discrimination of the items answered incorrectly in a testee's response pattern. This finding has been reported earlier by Samejima (1969) and indicates that the difficulty of items answered correctly or incorrectly has no effect on achievement level estimates obtained using the two-parameter M-L logistic scoring method.

Three-parameter model. The estimates of achievement level for each of the response patterns when a three-parameter item characteristic response model was assumed are shown in Table 4. It may be seen from this table that the M-L normal scoring algorithm failed to converge on an estimate for 7 of the 30 response patterns. The M-L logistic algorithm failed for 9 of the 30 patterns. These failures occurred when the likelihood function was too flat to allow the algorithm (a Newton-Raphson procedure; see Bejar & Weiss, 1979, pp. 10-11) to determine the point of maximization within 100 attempts. In this test the likelihood function was flattened because of the addition of the lower asymptote parameter, c, the "pseudo-guessing" parameter. The effect of this parameter is to lower the amount of information obtained from any single response, thereby flattening the likelihood function.

For both M-L scoring methods the nonconvergences occurred for the 6 response patterns which were given the lowest θ estimates by the Bayesian method (the value of -8.77 for the M-L normal method represents an artificial convergence). In addition, both M-L methods failed for the [0,1,0,1,1] response pattern, which represents the responses of an individual who answered easy items (Items 1 and 3) incorrectly and difficult items (Items 4 and 5) correctly. The M-L logistic scoring method also failed to converge for the [0,1,0,1,0] response pattern, in which incorrect responses were given to the items with lower discriminations and correct responses were given to the higher discriminating items. As Table 4 shows, because the Bayesian scoring method does not use an iterative procedure, θ estimates were obtained for all 30 response patterns.

Due to these convergence failures, it was appropriate to examine the differences in the three scoring methods' rankings by including in the rankings only those response patterns for which θ estimates were obtained by all three methods. These curtailed rankings are shown as Rank 2 in Table 4.

Table 4
Achievement Level Estimates and Rank Orders for Bayesian and Maximum-Likelihood (M-L) Scoring Methods Assuming a Three-Parameter ICC Response Model

Response Pattern[a]	Bayesian			M-L Normal			M-L Logistic		
	Estimate	Rank	Rank 2[b]	Estimate	Rank	Rank 2	Estimate	Rank	Rank 2
1,1,1,1,0	.91	1	1	1.58	1	1	1.56	1	1
1,1,0,1,1	.60	2	2	1.20	2	2	1.34	2	2
1,1,1,0,1	.53	3	3	.98	3	3	.89	4	4
1,1,1,0,0	.23	4	4	.37	5	5	.41	7	7
1,1,0,1,0	.16	5	5	.58	4	4	.58	5	5
0,1,1,1,1	.02	6	6	-.59	8	8	1.33	3	3
1,1,0,0,1	-.15	7	7	-.33	6	6	-.35	8	8
0,1,1,1,0	-.27	8	8	-.71	9	9	.51	6	6
1,1,0,0,0	-.33	9	9	-.47	7	7	-.49	9	9
1,0,1,1,1	-.33	10	10	-.96	12	12	-.99	12	12
0,1,1,0,1	-.49	11	11	-.77	10	10	-.57	10	10
1,0,1,1,0	-.53	12	12	-.99	13	13	-1.06	13	13
1,0,1,0,1	-.69	13	13	-1.01	14	14	-1.09	14	14
0,1,1,0,0	-.60	14	14	-.82	11	11	-.79	11	11
1,0,1,0,0	-.77	15	15	-1.03	15	15	-1.14	15	15
0,1,0,1,1	-.83	16	--	NC[c]	--	--	NC	--	--
0,1,0,1,0	-.92	17	--	-2.31	22	--	NC	--	--
0,1,0,0,1	-1.00	18	16	-1.45	16	16	-1.44	16	16
1,0,0,1,1	-1.04	19	17	-1.68	19[d]	19	-1.60	18	18
0,1,0,0,0	-1.05	20	18	-1.46	17	17	-1.50	17	17
1,0,0,1,0	-1.09	21	19	-1.68	19	19	-1.63	19.5	19.5
1,0,0,0,1	-1.15	22	20	-1.68	19	19	-1.63	19.5	19.5
1,0,0,0,0	-1.17	23	21	-1.69	21	21	-1.65	21	21
0,0,1,1,1	-1.31	24	--	NC	--	--	NC	--	--
0,0,1,1,0	-1.35	25	--	NC	--	--	NC	--	--
0,0,1,0,1	-1.39	26	--	NC	--	--	NC	--	--
0,0,1,0,0	-1.42	27	--	-8.77	23	--	NC	--	--
0,0,0,1,1	-1.70	28	--	NC	--	--	NC	--	--
0,0,0,1,0	-1.71	29	--	NC	--	--	NC	--	--
0,0,0,0,1	-1.72	30	--	NC	--	--	NC	--	--

[a]The response patterns [0,0,0,0,0] and [1,1,1,1,1] are not included because M-L estimates cannot be obtained for these response patterns.

[b]Ranking of response patterns for which all three methods obtained estimates.

[c]The M-L estimation algorithm failed to converge on a unique maximum.

[d]Ties were assigned the average of the ranks that the tied estimates would span if they were not tied.

Using these curtailed rankings, the Bayesian estimates differed in rank order from the M-L normal estimates for 15 of 21 response patterns. The average difference in rank position between the two methods was .95. The largest difference in ranks was 3. The Bayesian estimates also differed from the M-L logistic estimates for 14 of 21 response patterns. The average difference in ranks between these methods was .95 ranks, and the maximum difference was 3.

The M-L normal ranking differed from the M-L logistic ranking for 10 of 21 response patterns. The average difference between the rankings of the estimates derived from the two scoring method rankings was .81. The largest difference in rank order was 5.

The most obvious effect of the addition of the third parameter was that the achievement level estimates obtained by each of the three scoring methods were consistently lower than those obtained using the one- and two-parameter models. This result may be explained by the fact that the third parameter indicates the ease with which an item might be answered correctly without any knowledge of the subject matter. As the level of this parameter increases, the weight given to a correct answer is decreased for each of the scoring methods; therefore, the final θ estimates are lower.

For the response patterns for which each of the scoring methods obtained an achievement level estimate, the tendency for the Bayesian scoring method to result in more moderate estimates than either of the M-L methods was still evident, as it was under the one- and two-parameter models. Also, the tendency for the discrepancies between the estimates to be higher for response patterns in which the estimates were quite different from zero was still apparent, particularly in the comparison between the Bayesian method and the M-L normal method. For example, for the 3 response patterns giving rise to the most extreme θ estimates--[1,0,0,1,0], [1,0,0,0,1], and [1,0,0,0,0]--the average difference between the estimates was .55 score units; for the 3 response patterns for which the θ estimates were closest to zero--[1,1,0,1,0], [0,1,1,1,1], and [1,1,0,0,1]--the average difference between the estimates was .41 score units.

The M-L logistic estimates using the three-parameter model were not as obviously related to the discriminations of items answered incorrectly as in the two-parameter data. Thus, the three-parameter data permitted the first clear comparison of the differences between the Bayesian and M-L logistic estimates. In general, the Bayesian θ estimates tended to be less extreme (e.g., closer to zero) than the M-L logistic θ estimates, similar to the comparison between the Bayesian and M-L normal estimates. However, there was no trend for the estimates for the response patterns with extreme θ estimates to diverge to a greater extent than those with moderate θ estimates, as in the comparison between the Bayesian and M-L normal estimates.

Relationships among models and methods. Values of Kendall's Tau among achievement level estimates generated by the three scoring methods within each response model are shown in Table 5. The highest correlation between scoring methods was between the Bayesian method and the M-L normal method for both the one-parameter and two-parameter models (Tau=.963 and .948, respectively). For the three-parameter model, the most similar ranks were obtained by the two M-L methods (Tau=.918). For all three models, the least similar sets of rankings were derived from the Bayesian and M-L logistic methods. When the second and third parameters were added to the response models, there was a tendency for the correlations between pairs of scoring methods to become more similar as the correlations between the M-L logistic ranks and those of the other two scoring methods increased. At the same time, there was a decrease in the similarity of rankings produced by the Bayesian and M-L normal methods. Using the three-parameter model, the three pairs of correla-

tions tended to cluster around a Tau of .90, accounting for about 81% common variance in the pairs of rankings produced by the three scoring methods.

Table 5
Values of Kendall's Tau Among Achievement Estimates from Three Scoring Methods for Each ICC Response Model

	Response Model		
Scoring Methods	One-Parameter	Two-Parameter	Three-Parameter
Bayesian vs. M-L Normal	.963	.948	.906
Bayesian vs. M-L Logistic	.864	.873	.893
M-L Normal vs. M-L Logistic	.876	.898	.918

Conventional Test

Convergence failures. The data from the hypothetical test indicated that the M-L scoring methods failed to obtain achievement level estimates under certain circumstances. M-L scoring methods will be unable to converge for response patterns which include either all correct answers or all incorrect answers. In addition, there were other response patterns with likelihood functions that did not have a single obvious maximum. These kinds of response patterns will also result in convergence failures.

Table 6
Percentage of Maximum-Likelihood Convergence Failures for Conventional Test Data with Varying Numbers of Items (N=200)

	Percentage of Convergence Failures					
	One-parameter model		Two-parameter model		Three-parameter model	
Number of Items	M-L Normal	M-L Logistic	M-L Normal	M-L Logistic	M-L Normal	M-L Logistic
3	63	63	63	63	66	65
6	27	27	27	27	29	30
9	17	17	17	17	17	17
12	13	13	13	13	13	13
15	10	10	10	10	10	10
18	8	8	8	8	8	8
21	8	8	8	8	8	8
24	6	6	6	6	6	6
27	5	5	5	5	5	5
30	4	4	4	4	4	4
33	4	4	4	4	4	4
36	1	1	1	1	1	1
39	1	1	1	1	1	1

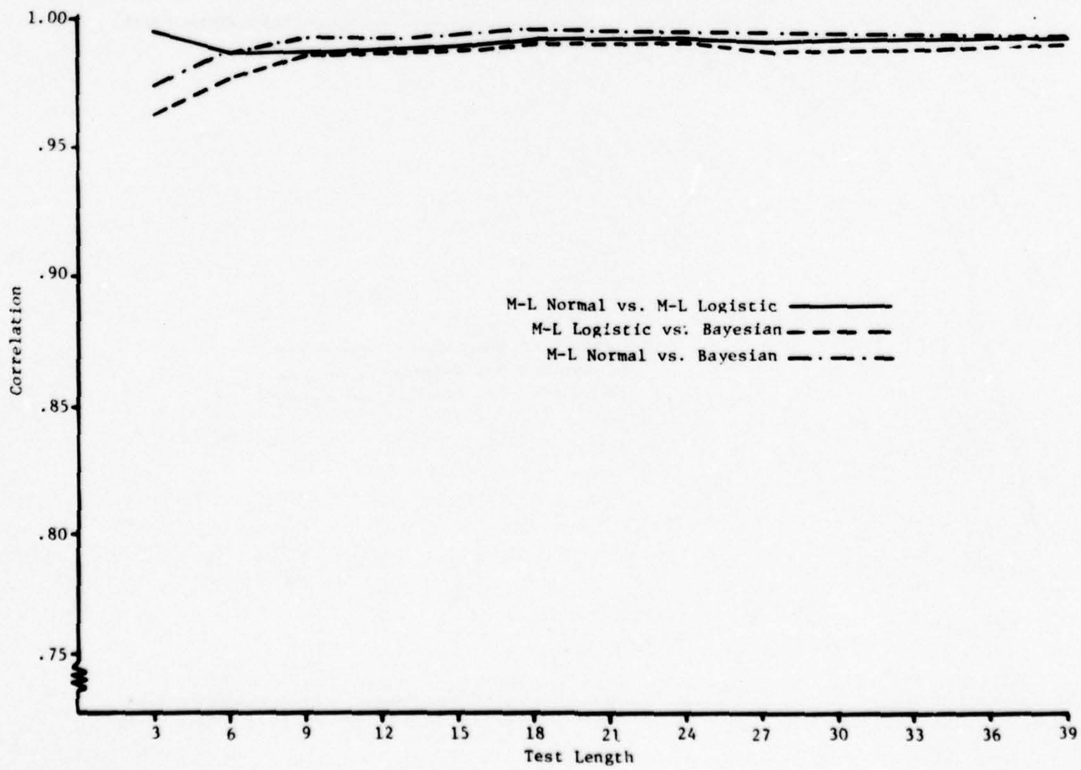
Table 6 shows the percentage of individuals for whom the M-L scoring methods did not converge on a unique achievement level estimate for each test

length and response model, using conventional test response data. The M-L scoring methods failed to obtain achievement level estimates for almost two-thirds of the response patterns at the shortest test length (3 items), regardless of the response model or the scoring method used. At a test length of 6 items, the convergence failure rate varied between 27% and 30% of the response patterns. For both 3-item and 6-item tests, there were no differences in the percentage of convergence failures between the M-L normal and M-L logistic scoring methods within the one-parameter and two-parameter models. Similarly, there were no differences between these two models regardless of scoring method. For both M-L logistic and M-L normal scoring methods, the three-parameter model resulted in slightly more convergence failures than the one- and two-parameter models, for 3- and 6-item tests.

For conventional tests of 9 or more items, there were no differences among models or methods of scoring in the rate of convergence failures. The percentage of convergence failures dropped consistently with increasing test length. But even for relatively long tests (e.g., 30 items), 4% of the 200 response patterns failed to converge within 100 iterations. At the longest test length (39 items), 1% of the response patterns failed to yield convergent estimates for all methods and models of M-L scoring.

One-parameter model. Appendix Table C shows Pearson product-moment correlations among scores derived from each pair of the three scoring methods for test lengths of 3 to 39 items, in steps of 3 items; these correlations were

Figure 1
Correlations Between Achievement Level Estimates as a Function of Test Length for Conventional Test Data Using a Two-Parameter Model

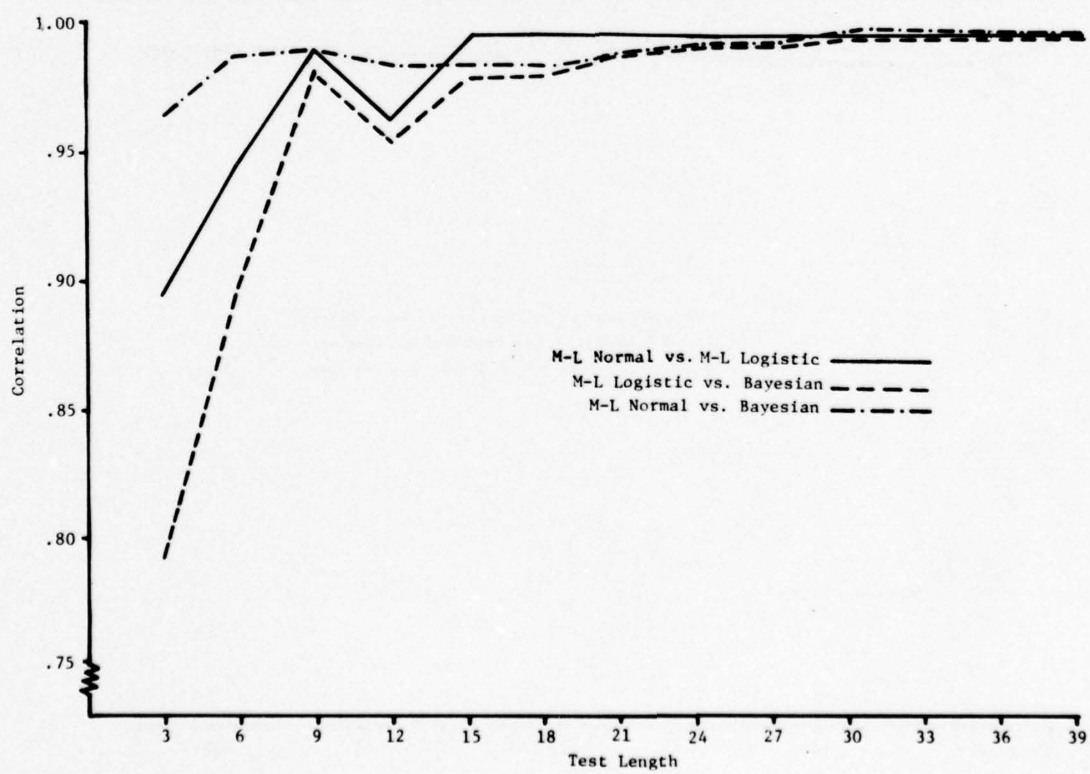


based on only those cases for which the M-L scoring estimates converged. As the data show, the minimum correlation was r=.9741 for scores from the M-L logistic and Bayesian methods for a 3-item test. The maximum r was .9967 for scores from the M-L normal and Bayesian methods for an 18-item test. There was no general trend in the data either as a function of test length or scoring method. In all cases, for tests greater than 3 items, more than 97% of the variance in a scoring method was common with the other scoring methods.

Two-parameter model. Figure 1 shows the correlations between scores derived from the three scoring methods when the data were scored by the two-parameter model (numerical values are in Appendix Table C). In general, the correlations were slightly lower than when the data were scored using only the difficulty parameter information. For the two-parameter data, the minimum correlation was .9629 between the M-L logistic and Bayesian methods, at a test length of 3 items. The highest correlation was .9958 between the M-L normal and M-L logistic methods for a 3-item test. As Figure 1 shows, there was a slight trend toward higher correlations as test length increased. For the two-parameter data, 97% of the variance in scores was common between all pairs of methods for test lengths greater than 6 items.

Three-parameter model. Figure 2 shows the correlations among the achievement level estimates obtained from each of the scoring methods at test lengths from 3 to 39 items when the data were scored using a three-parameter ICC response model (numerical values are in Appendix Table C). It can be seen

Figure 2
Correlations Between Achievement Level Estimates as a Function of Test Length for Conventional Test Data Using a Three-Parameter Model



from Figure 2 that the correlations among the three scoring methods were considerably lower for the three-parameter model at test lengths of 15 items or less than they were when only one or two parameters were used to score the data. The lowest correlation was r=.7917 for the M-L logistic versus Bayesian comparison for tests of 3 items; the highest correlation was r=.9967 for the M-L normal versus M-L logistic comparison for tests of 39 items. The lowest correlations occurred uniformly for 3-item tests, with large increases into the r=.90 range for all correlations for 6-item tests. There was a general trend for all correlations to increase with increasing test length, except for a slight drop at 12 items associated with the M-L logistic method. There were only very small differences among correlations at test lengths of 27 or more items. There was a general tendency throughout the data for scores from the M-L logistic and Bayesian methods to correlate lowest, with the trend most pronounced at shorter test lengths. For the three-parameter data, 97% of the variance in each scoring method was common with the other scoring methods for tests 15 items or more in length.

Summary. The data show a general decrease in similarity among scores as more parameters were used to score the items. The addition of the discrimination parameter tended to reduce correlations among scoring methods slightly for tests of less than 9 items in length; however, there were no large differences between scoring methods for the two-parameter data. When the "guessing" parameter was added, there was a marked decrease in similarity among scores associated with the M-L logistic method for tests shorter than 18 items; relationships between the M-L normal scores and the Bayesian scores remained high, although they were somewhat lower for most test lengths than with two-parameter scoring.

Adaptive Test

Convergence failures. Table 7 shows the percentage of response patterns for which the M-L scoring methods failed to obtain an achievement level estimate at each test length from 3 to 48 items using each response model. These data show that there were no consistent differences between the M-L logistic and M-L normal scoring methods and no differences at all between these methods using the one- and two-parameter response models.

Under each response model, 20 to 38% of the response patterns resulted in estimation failures for the shortest test length. Fewer estimation failures were noted at longer test lengths. For the one- and two-parameter models, no convergence failures were observed for any test length greater than 9 items. Under the assumption of the three-parameter model, more convergence failures were noted than for the simpler response models for test lengths up to 33 items. No convergence failures were observed at any test length greater than 33 items.

These results were not completely comparable to convergence failures observed for the conventional test because of the stradaptive variable length termination. At longer test lengths the number of testees on which the percentages were based dropped steadily as the ceiling stratum for individuals was determined. This variable termination criterion may add an unknown amount of bias to comparisons made between the conventional and adaptive tests in this study.

Table 7
Percentage of Maximum Likelihood Convergence Failures for Adaptive Test Data with Varying Numbers of Items

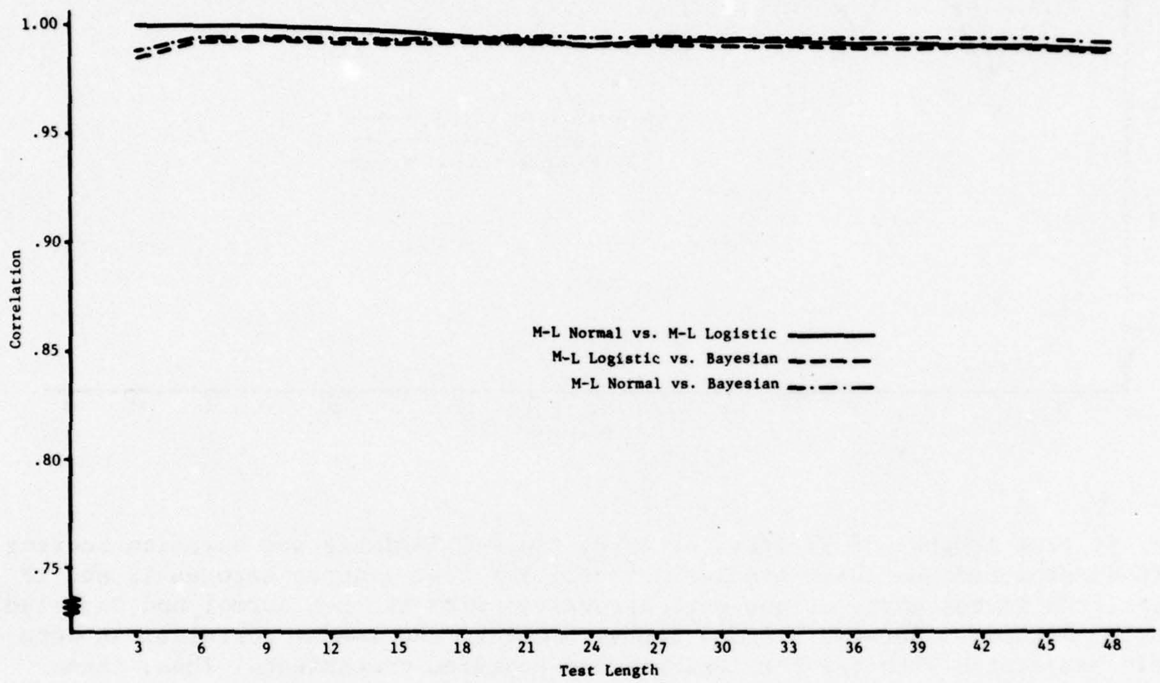
Number of Items	Number of Individuals	Percentage of Convergence Failures					
		One-parameter model		Two-parameter model		Three-parameter model	
		M-L Normal	M-L Logistic	M-L Normal	M-L Logistic	M-L Normal	M-L Logistic
3	200	20	20	20	20	38	30
6	200	6	6	6	6	9	11
9	200	1	1	1	1	4	6
12	185	0	0	0	0	1	2
15	169	0	0	0	0	1	2
18	143	0	0	0	0	1	2
21	127	0	0	0	0	2	3
24	108	0	0	0	0	0	0
27	97	0	0	0	0	1	1
30	83	0	0	0	0	2	1
33	79	0	0	0	0	1	0
36	67	0	0	0	0	0	0
39	60	0	0	0	0	0	0
42	56	0	0	0	0	0	0
45	51	0	0	0	0	0	0
48	47	0	0	0	0	0	0

One-parameter model. Appendix Table D shows Pearson product-moment correlations between achievement level estimates derived from each pair of the three scoring methods for test lengths of 3 to 48 items. These correlations were based only on those individuals for whom the M-L scoring methods did not fail to converge and for whom the test continued to the specified test length. The data show that the lowest observed correlation was .9927 for scores from the M-L logistic and Bayesian methods for a test length of 3 items. The highest observed correlation was .9998, between scores from the M-L logistic and M-L normal methods at the 9-item test length and from the M-L normal and Bayesian methods at all test lengths between 24 and 45 items. For all test lengths, more than 97% of the score variance for each scoring method was common with every other scoring method.

Two-parameter model. Figure 3 shows the correlations between achievement level estimates derived from each pair of the three scoring methods as a function of test length, assuming a two-parameter response model (numerical values are shown in Appendix Table D). These correlations were, in general, slightly lower than those observed under the one-parameter model. The lowest observed correlation was .9854, between scores obtained from the M-L logistic and Bayesian methods for a test length of 3 items. The highest observed correlation was .9996, between scores from the M-L logistic and M-L normal methods, also at a test length of 3 items. Again, at all test lengths, more than 97% of the score variance in a scoring method was common with every other method. As with the one-parameter model, no general trend was noted in the data as a function of test length, other than a very slight tendency for the

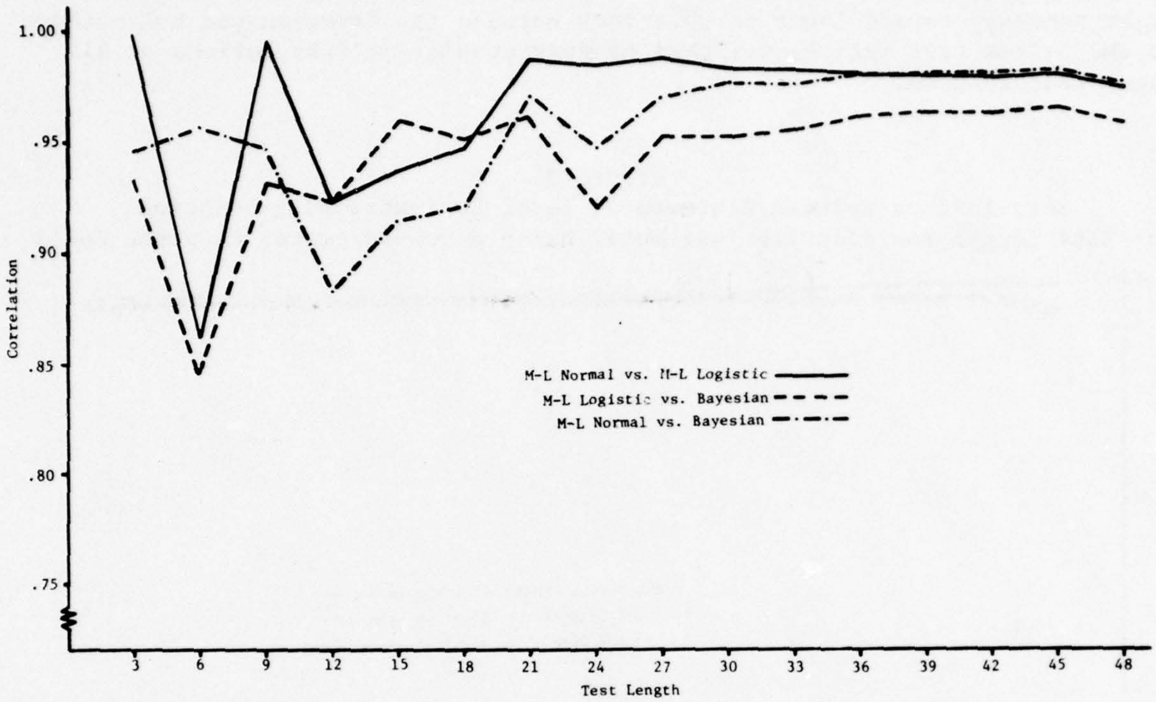
correlation between scores from the M-L normal and M-L logistic methods to decrease as the test length increased; but even at the longest test length observed (48 items), this correlation was still .9892. Figure 3 shows a slight tendency toward lower correlations between the Bayesian and M-L methods for the 3-item test length, followed by very consistent correlations at all longer test lengths.

Figure 3
Correlations Between Achievement Level Estimates as a Function of Test Length for Adaptive Test Data Using a Two-Parameter Response Model



Three-parameter model. Figure 4 shows the correlations between scores obtained from each pair of the three scoring methods as a function of test length for the three-parameter model (numerical values are in Appendix Table D). It is evident from this figure that the very consistent and high correlations observed under the assumption of the one- and two-parameter models were not observed when the three-parameter model was assumed, particularly for shorter test lengths. The lowest correlation observed under the assumption of the three-parameter model was .8444, between scores from the M-L logistic and Bayesian models at the 6-item test length. The highest correlation observed was .9997, between estimates from the M-L logistic and M-L normal methods at the 3-item test length. There was a general tendency for the correlations among the scores obtained from each pair of the three scoring methods to become higher and more consistent at longer test lengths. There was, however, no test length for which more than 97% of the score variance was common among the three scoring methods. This is the only combination of testing method and response model examined in this study for which this common variance criterion was not met at any test length.

Figure 4
Correlations Between Achievement Level Estimates as a Function of Test Length for Adaptive Test Data Using a Three-Parameter Model



At test lengths of 21 items or more, the M-L logistic and Bayesian scoring methods produced the least similar scores. For test lengths between 12 and 18 items, the lowest correlations were associated with the M-L normal and Bayesian scoring methods. Between 3 and 9 items, however, the lowest correlations were again associated with the M-L logistic and Bayesian comparison. Thus, these data show a general tendency for the Bayesian θ estimates to be consistently less similar to the M-L estimates than were the θ estimates for the two M-L scoring methods.

Summary. These data show a tendency toward greater dissimilarity among scores obtained from the three scoring methods when more complex response models were used to score the item responses from the adaptive test data. The use of a varying discrimination parameter in the two-parameter model reduced all observed correlations slightly (.0062 on the average), and the correlations between M-L logistic scores and Bayesian scores most noticeably (.0073 on the average). When a nonzero "guessing" parameter was used in the three-parameter model to obtain achievement level estimates, correlations among scores from the three different scoring methods decreased to a much greater extent (.0350 mean decrease), with the greatest decrease again being observed in correlations between scores from the M-L logistic and Bayesian methods (.0460 mean decrease). The three-parameter results showed less similarity among the scores obtained from the three scoring methods than either the one- or two-parameter results for each test length; differences among the achievement level estimates for

the one- and two-parameter models might be called unimportant, since correlations between the estimates were consistent for tests of reasonable lengths and tended to differ very little from 1.0. The three-parameter response model yielded consistently lower correlations between scores obtained using the three scoring methods; these correlations did not approach 1.0, even for long test lengths.

Comparison of Conventional and Adaptive Data

For the one-parameter model, correlations between scores obtained through the three different scoring methods were uniformly high; but those obtained from the adaptive testing procedure tended to be slightly higher than those obtained from the conventional testing procedure, for all test lengths. Using the one-parameter model with conventional test data, the average correlation observed between scores obtained from all pairs of scoring methods across all test lengths was .9920; for the adaptive test data, the average correlation was .9990.

Under the assumption of the two-parameter model, there was still a trend for the correlations between scores to be higher for data from the adaptive testing procedure than for data from the conventional testing procedure; but this trend was not as strong as that observed under the assumption of the one-parameter model. For the two-parameter model, the average observed correlation between scores from the three scoring methods across all test lengths for the conventional test was .9900. For the adaptive test data, the average correlation was .9929.

Under the assumption of the three-parameter model, the mean correlation between scores from the three scoring procedures for all test lengths was .9799 using responses to the conventional test and .9582 using responses to the adaptive test. Under this response model, the trend was for the scores obtained from the conventional test to be more consistent across the three scoring models than the scores obtained from the adaptive test. This trend is the opposite of the trend observed for the one- and two-parameter models.

One further point is of interest for the comparison of the adaptive and conventional testing procedures. Tables 6 and 7 show that the adaptive test data resulted in fewer M-L convergence failures than the conventional test data at every comparable test length. This difference resulted in 40% to 100% fewer observed estimation failures for the adaptive testing procedure. For the one- and two-parameter models, no estimation failures were observed at any test length greater than 9 items for the adaptive test data; for the conventional test data, estimation failures were observed at every test length up to 39 items, the longest test length examined. Using the three-parameter model, no estimation failures were observed at any test length greater than 33 items for the adaptive test data; but failures were observed for the conventional data up to the longest test length of 39 items.

Discussion and Conclusions

The data show that under certain conditions, the three ICC-based scoring methods will result in different achievement level estimates. Trends evident in the hypothetical test data were, in some cases, clarified by the analysis

of the conventional and adaptive test data. The data from the hypothetical five-item test clearly illustrated that θ estimates from the one-parameter logistic model scored by maximum likelihood are directly related to the number of items answered correctly, regardless of the difficulties of the items answered correctly or incorrectly. It is this property of the one-parameter logistic model which permits the Rasch model to use the number-correct score within an ICC framework. When all three scoring methods were applied to the same data, however, the results indicated that the M-L logistic scoring method in the one-parameter case ignored information that allowed differentiation among dissimilar response patterns having the same number-correct score. From an ICC point of view, promising fuller use of test response information, the one-parameter M-L logistic scoring method is no more informative than the number-correct score which it reflects, at least for short tests similar to the five-item hypothetical test. When the three scoring models were applied to live-testing data from both conventional and adaptive tests, correlations among θ estimates derived from the one-parameter model were quite high, regardless of test length. Thus, in the live-testing data, the fact that the M-L logistic scoring method ignored the item difficulties did not seriously affect its performance in comparison to the other two scoring methods.

When the hypothetical test data were scored using both the difficulty and discrimination parameters, the M-L logistic method still did not use the item difficulties in arriving at θ estimates. In this case, the M-L logistic θ estimates were associated, not with number-correct scores, but with the item discriminations; individuals who incorrectly answered items of the same discrimination, but with differing difficulties, all received the same θ estimate. Again, both the Bayesian and M-L normal scoring methods provided differential and highly correlated θ estimates, which took into account both the response pattern data and the item difficulties and discriminations. In live-testing data, in which all possible response patterns are unlikely to occur (as they did in the hypothetical test data), this trend again seemed to lack practical importance. In both the adaptive and conventional test data scored by the two-parameter model, correlations among θ estimates were very high, regardless of test length.

Both the one-and two-parameter hypothetical data illustrated the tendency of the Bayesian θ estimates to be regressed toward the mean. That is, the Bayesian scoring method provided lower θ estimates for scores above the mean and higher θ estimates for scores below the mean, in comparison to the two M-L scoring methods. This trend continued in the three-parameter data, although both rank-order and product-moment correlations remained high, as in the former two analyses. This result, however, has implications for the use of the Bayesian scoring method in any applied situation in which the absolute, as opposed to relative, level of the θ estimates is of importance. Since the Bayesian scoring method tends to restrict the range of θ estimates by imposing a normal distribution on them, θ estimates beyond ± 2.0 will rarely be obtained. The result is likely to be a tendency for this scoring method to fail to identify and/or to distinguish accurately among testees with extreme θ estimates.

The dissimilarities among the three scoring methods became most evident when the data were scored using the three-parameter model. The major dissimilarity, evident in all three data sets, was between the Bayesian and M-L logistic methods. In the adaptive test data, the Bayesian scoring method produced θ

estimates which had lowest correlations with one of the two M-L methods at all test lengths. For conventional tests of less than 15 items and for adaptive tests at all the lengths used in this study, these differences were substantial, indicating markedly different orderings of individuals, as in the hypothetical test data.

The three-parameter data also illustrated two other trends. First, the hypothetical test data showed a tendency toward lower θ estimates when the c parameter was included in scoring. A second, and more practically troublesome, trend was the tendency toward more convergence failures with the three-parameter data. This result was obvious in both the hypothetical test data and the live-testing data. The tendency toward convergence failures for the M-L scoring methods was most obvious in the conventional test; the number of convergence failures in the adaptive test was considerably less than in the conventional test when number of items was equal. This occurred because adaptive tests tend to locate for each testee the region of the item pool in which the testee will answer about half of the items correctly and half incorrectly. Thus, except for the rare individual for whom the adaptive test item pool is completely inappropriate in difficulty, adaptive tests will result in response patterns that are more likely scorable by M-L methods. This is not true of fixed-item peaked conventional tests, which must be targeted for a specific population θ level and which may be too easy or too difficult for substantial numbers of testees, resulting in response patterns not scorable by M-L methods.

Choosing a Scoring Method

These data show that in an adaptive test or in a situation in which a short conventional test is being administered, the choice of one of the ICC-based methods over another may have an impact on the ranking of the students in a course of training. For these situations, it is important that educators choose a scoring method most aligned to their philosophy of grading. To determine the "correct" scoring method to use, the underlying philosophies of the different scoring methods may be viewed by examining the relationship of the scores obtained from a particular method to the ICC response model underlying the test.

This can be illustrated with the hypothetical test used in the example of the two-parameter model, which was borrowed, in part, from Samejima (1969). Because the item parameters for this test were known, the way in which each scoring method depends on the item difficulty and discrimination parameters of the items answered by the testees may be examined. From inspection of Table 3 for the two-parameter data, it can be seen that the Bayesian strategy gave results most similar to a number-correct scoring strategy, since it ordered individuals almost perfectly with respect to number correct. However, higher rankings resulted with the Bayesian scoring method for individuals correctly answering more difficult (high b) and more discriminating (high a) items. A disadvantage of this scoring approach, however, is that more weight is given to the early items in the test.

The M-L normal rankings can be characterized as being dependent upon both the a and b parameters, but the dependence is less easily described than that of the Bayesian strategy. The M-L normal estimates tended to reward

correct answers to difficult items or correct answers to more discriminating items and to penalize inconsistent response patterns (that is, incorrect answers to easy items and correct answers to difficult items). The M-L logistic rankings for this response model were independent of the difficulty of the items answered correctly or incorrectly. As pointed out earlier, rankings were totally dependent on the discriminatory power of the items answered incorrectly by the individual (see Samejima, 1969, for the theoretical rationale).

It appears, therefore, that under the two-parameter response model, the M-L normal scoring method allows the most freedom from number-correct scoring and makes the most use of the parameter values of the items. If educators feel that this "philosophy" is in accord with their own, then it is the one that should be used; if it is not, one of the other scoring methods may serve better.

In addition to this "philosophy of scoring" approach, some of the other characteristics of the scoring methods should be considered. For instance, the Bayesian method allows the use of prior information in obtaining an achievement level estimate. If this prior information is accurate, this might be an advantage for obtaining good θ estimates from a short test. Prior information is not useful for M-L estimation. But if available prior information is not correct, the M-L scoring methods will be more accurate than the Bayesian method.

One final difference between the Bayesian and M-L scoring methods may be of some importance to educators. When individuals are able to answer test questions correctly by guessing, as in a multiple-choice test, the three-parameter ICC response model is most appropriate for scoring the test responses. Using this response model, M-L scoring methods will fail to converge on a unique θ estimate in some cases. For conventional test response data (Table 6), the percentage of such failures remained rather high under both M-L scoring methods (at least 5%) until more than 27 items had been administered. At no test length did all cases converge in the conventional test data.

The adaptive testing procedure fared better in this respect (Table 7). After the adaptive administration of only 9 items, neither M-L scoring method failed to obtain θ estimates in more than 3% of the cases. Further, all response patterns resulted in convergent θ estimates at all test lengths greater than 33 items.

These results suggest that an educator might take two courses of action to avoid the estimation failures of M-L scoring methods. One approach is to use a Bayesian scoring method, but with cognizance of its tendency to regress all θ estimates toward the mean. The other solution, of course, is to use an adaptive testing procedure in conjunction with either M-L scoring method.

In the final analysis, however, the choice of scoring method should be based on the validity of scoring methods in the prediction of external criteria. This study has demonstrated that, at least under the three-parameter ICC model, different scoring methods will provide different θ estimates. Given this knowledge, the question becomes one of studying the validity of the scores obtained from the different scoring methods with respect to relevant external criteria in order to determine whether the observed differences result in the differential predictability of criterion performance.

References

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, January 1979.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495).

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828).

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062).

Lindgren, Bernard W. Statistical theory. New York: Macmillan, 1976.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964).

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielson & Lydiche, 1960.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study (Research Report 77-1). Columbia: University of Missouri, Educational Psychology Department, Tailored Testing Research Laboratory, November 1977.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34 (4, Pt. 2, Monograph No. 17).

Sympson, J. B. Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1977. (NTIS No. AD A038114).

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1975, pp. 97-102. (Superintendent of Documents Stock No. 006-00940-9).

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270).

Appendix: Supplementary Tables

Table A
Parameter Estimates for Items in the Conventional Test

Item No.	No. Testees	*a*	*b*	*c*
3060	1323	.86	-1.31	.29
3067	1217	1.07	-.76	.21
3065	1324	1.17	-1.66	.39
3056	1134	.71	.89	.26
3063	1084	.91	1.51	.37
3073	1314	1.43	-1.57	.31
3058	1283	1.05	-.43	.44
3274	1274	.85	-1.05	.26
3271	1166	.95	1.32	.30
3055	1265	1.71	-.65	.24
3072	1177	1.02	.65	.32
3057	1285	1.20	-1.35	.26
3064	1287	.94	.86	.24
3069	1247	.88	-.01	.48
3054	1258	1.29	-.93	.31
3066	1057	1.05	.53	.31
3268	1211	.97	-.28	.18
3267	1285	1.02	-1.22	.23
3272	1274	1.06	-.81	.37
3070	1252	.95	-1.28	.22
3008	891	.96	-1.75	.18
3019	782	1.31	.29	.29
3062	1215	1.47	.43	.30
3061	1078	.85	1.57	.30
3262	1275	.81	.47	.45
3263	1092	.99	2.29	.53
3447	1266	1.18	.93	.32
3443	1264	1.07	-1.64	.37
3438	1095	.70	.21	.27
3448	1294	1.40	.73	.30
3435	1258	.83	-.61	.42
3439	1091	1.36	.64	.32
3436	1018	1.12	1.59	.41
3449	1138	.91	1.26	.14
3440	957	1.52	2.00	.30
3437	1147	1.95	.66	.28
3427	773	.92	1.51	.26
3445	1282	1.19	.44	.34
3444	1139	.88	.78	.38

Table B

Item Number, Number of Testees in Parameterization Group, Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Items in the Stradaptive Item Pool

Item	*N*	*a*	*b*	*c*	Item	*N*	*a*	*b*	*c*	Item	*N*	*a*	*b*	*c*
Stratum 9 (15 items)					Stratum 6 (19 items)					Stratum 3, cont.				
3209	740	2.50	2.29	.29	3047	608	1.66	.44	.29	3011	864	1.32	-.86	.20
3417	539	2.50	3.00	.35	3079	952	1.61	.27	.35	3435	1258	.83	-.61	.35
3033	328	1.54	2.44	.35	3213	900	.93	.52	.35	3216	809	1.27	-.62	.18
3440	957	1.52	2.00	.30	3041	716	1.51	.23	.35	3054	1258	1.29	-.93	.31
3251	523	2.50	2.39	.35	3062	1215	1.47	.43	.30	3221	938	1.25	-.52	.17
3406	519	1.31	2.48	.35	3405	770	1.40	.55	.32	3049	814	1.15	-.71	.18
3045	680	1.02	2.48	.27	3445	1282	1.19	.44	.34	3255	657	1.14	-.72	.26
3242	613	.94	2.40	.35	3218	500	.82	.58	.12	3067	1217	1.07	-.76	.21
3407	564	1.02	2.41	.29	3019	782	1.31	.29	.29	3246	656	1.10	-.72	.28
3263	1092	.99	2.29	.35	3207	915	.70	.46	.28	3022	620	1.01	-.48	.30
3241	756	.91	2.09	.17	3431	780	.70	.28	.34	3272	1274	1.06	-.81	.35
3414	368	.88	2.29	.32	3000	844	1.24	.52	.35	3017	950	.99	-.58	.16
3402	401	.83	2.44	.35	3046	626	1.18	.24	.22	3076	1054	.94	-.73	.21
3247	718	.82	2.42	.35	3042	626	1.15	.37	.27	3224	869	.80	-.50	.37
3228	396	.67	2.49	.31	3050	713	1.13	.35	.18	Mean		1.22	-.68	.22
Mean		1.33	2.39	.32	3066	1057	1.05	.53	.31					
					3034	639	1.01	.37	.28	Stratum 2 (20 items)				
Stratum 8 (20 items)					3262	1275	.81	.47	.35	3023	667	2.40	-1.15	.35
3409	602	2.50	1.28	.00	3438	1095	.70	.21	.27	3202	922	1.81	-.99	.21
3234	220	2.50	1.73	.00	Mean		1.14	.40	.29	3415	915	.85	-.96	.35
3018	953	.89	1.25	.35						3245	885	1.34	-.96	.21
3204	505	1.14	1.66	.35	Stratum 5 (15 items)					3236	667	1.26	-1.20	.33
3422	589	1.47	1.50	.35	3282	1037	2.06	-.02	.35	3020	915	1.23	-1.28	.17
3411	767	1.36	1.23	.35	3220	896	1.79	-.03	.26	3028	677	1.12	-1.26	.35
3250	373	.91	1.94	.29	3005	831	1.43	.11	.35	3226	941	1.09	-.98	.20
3206	410	.74	1.51	.21	3425	649	1.36	.17	.23	3210	895	1.04	-1.22	.35
3410	427	1.30	1.34	.31	3039	908	1.12	.12	.00	3239	960	1.04	-1.13	.21
3429	780	1.25	1.24	.28	3214	809	1.12	.03	.23	3013	880	1.00	-.97	.35
3419	342	1.23	1.48	.25	3412	664	1.12	.19	.35	3267	1285	1.02	-1.22	.23
3421	750	1.17	1.15	.35	3051	752	1.29	.21	.28	3257	928	.98	-1.02	.25
3436	1018	1.12	1.59	.35	3279	969	.99	.01	.28	3070	1252	.95	-1.28	.22
3271	1166	.95	1.32	.30	3403	626	.99	.18	.19	3036	872	.92	-1.18	.16
3061	1078	.95	1.57	.30	3069	1247	.88	-.01	.35	3014	907	.86	-1.24	.14
3427	773	.92	1.51	.26	3211	628	.88	.01	.13	3060	1323	.86	-1.31	.29
3449	1138	.91	1.26	.14	3002	929	.82	.13	.14	3274	1274	.85	-1.05	.26
3063	1084	.91	1.51	.35	3426	870	.68	.07	.22	3238	837	.82	-1.06	.21
3074	671	.84	1.79	.35	3423	682	.66	.16	.27	3032	857	.77	-1.06	.27
3420	541	.68	1.62	.35	Mean		1.15	.09	.24	Mean		1.11	-1.13	.26
Stratum 7 (20 items)					Stratum 4 (13 items)					Stratum 1 (17 items)				
3408	451	2.50	1.05	.31	3256	649	2.31	-.33	.26	3077	1053	2.50	-1.39	.20
3437	1147	1.95	.66	.28	3430	903	1.15	-.30	.29	3027	667	1.67	-1.38	.35
3258	911	1.24	.81	.35	3031	851	1.47	-.33	.35	3443	1264	1.07	-1.64	.35
3432	595	1.72	.67	.35	3254	653	3.38	-.17	.22	3249	910	.91	-1.69	.17
3048	589	1.35	.66	.33	3237	895	1.54	-.37	.18	3428	899	.90	-1.56	.35
3413	832	1.40	.76	.35	3404	897	.65	-.29	.35	3073	1314	1.43	-1.57	.31
3448	1294	1.40	.73	.30	3244	854	1.35	-.44	.23	3205	908	1.25	-1.53	.19
3439	1091	1.36	.64	.32	3058	1283	1.05	-.43	.35	3078	1060	1.24	-1.65	.35
3219	520	1.23	.62	.21	3240	702	.98	-.28	.15	3057	1285	1.20	-1.35	.26
3072	1177	1.02	.65	.32	3268	1211	.97	-.28	.18	3065	1324	1.17	-1.66	.35
3277	892	1.00	1.04	.35	3208	850	.76	-.16	.12	3235	906	1.15	-1.40	.28
3035	772	.90	.68	.28	3006	676	.77	-.37	.33	3029	957	1.13	-1.50	.28
3433	657	1.35	.86	.30	3259	879	.69	-.41	.20	3201	902	1.07	-1.34	.23
3447	1266	1.18	.93	.32	Mean		1.23	-.32	.25	3008	891	.96	-1.75	.18
3064	1287	.94	.86	.24						3252	898	.79	-1.77	.35
3230	895	.90	.87	.35	Stratum 3 (19 items)					3003	914	.96	-1.76	.34
3444	1139	.88	.78	.35	3021	906	1.96	-.49	.21	3044	913	.87	-1.42	.15
3012	653	.75	.80	.35	3217	893	1.06	-.48	.14	Mean		1.19	-1.55	.28
3260	877	.71	.84	.28	3038	951	1.71	-.93	.00					
3056	1139	.71	.89	.26	3055	1265	1.71	-.65	.24					
Mean		1.22	.79	.31	3215	887	1.59	-.82	.23					

Table C

Correlations Between Achievement Level Estimates from Three Scoring Methods at Various Test Lengths for Conventional Test Data Scored by One-, Two- and Three-Parameter Models

(N=200[a])

Number of Items	One-Parameter Model N	One-Parameter Model MLL vs. MLN	One-Parameter Model MLL vs. Bayes	One-Parameter Model MLN vs. Bayes	Two-Parameter Model N	Two-Parameter Model MLL vs. MLN	Two-Parameter Model MLL vs. Bayes	Two-Parameter Model MLN vs. Bayes	Three-Parameter Model MLL vs. MLN N	Three-Parameter Model MLL vs. MLN r	Three-Parameter Model MLL vs. Bayes N	Three-Parameter Model MLL vs. Bayes r	Three-Parameter Model MLN vs. Bayes N	Three-Parameter Model MLN vs. Bayes r
3	75	.9955	.9741	.9832	75	.9958	.9629	.9749	70	.8957	71	.7917	70	.9852
6	146	.9831	.9867	.9964	146	.9874	.9776	.9876	140	.9463	141	.8994	142	.9881
9	167	.9847	.9892	.9957	167	.9864	.9858	.9921	167	.9890	167	.9797	167	.9892
12	174	.9858	.9888	.9959	174	.9868	.9859	.9920	174	.9619	174	.9527	174	.9828
15	181	.9893	.9904	.9963	181	.9889	.9881	.9939	181	.9953	181	.9792	181	.9846
18	184	.9920	.9926	.9967	184	.9918	.9910	.9951	184	.9960	184	.9800	184	.9822
21	184	.9912	.9920	.9966	184	.9916	.9912	.9947	184	.9950	184	.9877	184	.9882
24	188	.9921	.9935	.9960	188	.9924	.9922	.9947	188	.9953	188	.9901	188	.9905
27	191	.9918	.9905	.9957	191	.9919	.9890	.9944	191	.9953	191	.9900	191	.9909
30	192	.9928	.9912	.9959	192	.9927	.9897	.9950	192	.9953	192	.9932	192	.9961
33	192	.9935	.9915	.9961	192	.9934	.9897	.9952	192	.9957	192	.9938	192	.9963
36	198	.9946	.9925	.9958	198	.9945	.9907	.9948	198	.9961	198	.9943	198	.9954
39	198	.9954	.9928	.9958	198	.9951	.9911	.9948	198	.9967	198	.9948	198	.9960

[a]Differences in N from 200 represent nonconvergent cases in M-L scoring.

Table D
Correlations Between Achievement Level Estimates from Three Scoring Methods at Various Test Lengths for Adaptive Test Data Scored by One-, Two-, and Three-Parameter Models

Number of Items	One-Parameter Model						Two-Parameter Model						Three-Parameter Model					
	MLL vs. MLN		MLL vs. Bayes		MLN vs. Bayes		MLL vs. MLN		MLL vs. Bayes		MLN vs. Bayes		MLL vs. MLN		MLL vs. Bayes		MLN vs. Bayes	
	N	*r*	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*
3	159	.9992	159	.9927	159	.9960	159	.9854	159	.9854	159	.9871	108	.9997	141	.9334	124	.9456
6	188	.9995	188	.9978	188	.9988	188	.9923	188	.9923	188	.9933	174	.8628	177	.8444	182	.9563
9	198	.9998	198	.9986	198	.9991	198	.9946	198	.9946	198	.9950	189	.9970	189	.9315	192	.9479
12	185	.9997	185	.9984	185	.9989	185	.9920	185	.9920	188	.9937	182	.9228	182	.9243	184	.8832
15	169	.9997	169	.9989	169	.9995	169	.9923	169	.9923	169	.9941	166	.9382	166	.9599	168	.9154
18	143	.9997	143	.9990	143	.9995	143	.9930	143	.9930	143	.9944	150	.9478	140	.9520	142	.9226
21	127	.9995	127	.9989	127	.9996	127	.9924	127	.9924	127	.9946	123	.9883	123	.9617	125	.9711
24	108	.9995	108	.9990	108	.9998	108	.9915	108	.9915	108	.9946	108	.9846	108	.9201	108	.9485
27	97	.9995	97	.9990	97	.9998	97	.9914	97	.9914	97	.9951	96	.9880	96	.9534	96	.9702
30	83	.9995	83	.9990	83	.9998	83	.9909	93	.9909	83	.9944	81	.9830	82	.9536	81	.9769
33	75	.9994	75	.9990	75	.9998	75	.9910	75	.9910	75	.9946	74	.9827	75	.9566	74	.9766
36	67	.9994	67	.9989	67	.9998	67	.9900	67	.9900	67	.9948	67	.9810	67	.9624	67	.9815
39	60	.9993	59	.9989	60	.9998	60	.9900	60	.9900	60	.9946	60	.9804	60	.9648	60	.9822
42	56	.9995	45	.9990	56	.9998	56	.9907	56	.9807	56	.9946	56	.9800	56	.9656	56	.9823
45	51	.9994	51	.9990	51	.9998	51	.9906	51	.9906	51	.9946	51	.9806	51	.9669	51	.9828
48	47	.9994	47	.9988	47	.9997	47	.9892	47	.9892	47	.9933	47	.9765	47	.9598	47	.9784

DISTRIBUTION LIST

Navy

1 Dr. Ed Aiken
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-71
NAVTRAEQUIPCEN
Orlando, FL 32813

1 MR. MAURICE CALLAHAN
Pers 23a
Bureau of Naval Personnel
Washington, DC 20370

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Dr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 CAPT. D.M. GRAGG, MC, USN
HEAD, SECTION ON MEDICAL EDUCATION
UNIFORMED SERVICES UNIV. OF THE
HEALTH SCIENCES
6917 ARLINGTON ROAD
BETHESDA, MD 20014

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 CHAIRMAN, LEADERSHIP & LAW DEPT.
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMYY
ANNAPOLIS, MD 21402

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 CAPT Richard L. Martin
USS Francis Marion (LPA-249)
FPO New York, NY 09501

1 Dr. James McBride
Code 301
Navy Personnel R&D Center
San Diego, CA 92152

2 Dr. James McGrath
Navy Personnel R&D Center
Code 306
San Diego, CA 92152

1 DR. WILLIAM MONTAGUE
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Commanding Officer
Naval Health Research
Center
Attn: Library
San Diego, CA 92152

1 Naval Medical R&D Command
Code 44
National Naval Medical Center
Bethesda, MD 20014

1 Library
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 OFFICE OF CIVILIAN PERSONNEL
(CODE 26)
DEPT. OF THE NAVY
WASHINGTON, DC 20390

1 JOHN OLSEN
CHIEF OF NAVAL EDUCATION &
TRAINING SUPPORT
PENSACOLA, FL 32509

1 Psychologist
ONR Branch Office
495 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Code 436
Office of Naval Research
Arlington, VA 22217

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
OFFICE OF NAVAL RESEARCH BRANCH
223 OLD MARYLEBONE ROAD
LONDON, NW, 15TH ENGLAND

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco, CA 96503

1 Head, Research, Development, and Studies
(OP102X)
Office of the Chief of Naval Operations
Washington, DC 20370

1 Scientific Advisor to the Chief of
Naval Personnel (Pers-Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington, DC 20370

1 DR. RICHARD A. POLLAK
ACADEMIC COMPUTING CENTER
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 Mr. Arnold Rubenstein
Naval Personnel Support Technology
Naval Material Command (08T244)
Room 1044, Crystal Plaza #5
2221 Jefferson Davis Highway
Arlington, VA 20360

1 A. A. SJOHOLM
TECH. SUPPORT, CODE 201
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr. Robert Smith
Office of Chief of Naval Operations
OP-987E
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 CDR Charles J. Theisen, JR. MSC, USN
Head Human Factors Engineering Div.
Naval Air Development Center
Warminster, PA 18974

1 W. Gary Thomson
Naval Ocean Systems Center
Code 7132
San Diego, CA 92152

1 Dr. Ronald Weitzman
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 HQ USAREUE & 7th Army
ODCSOPS
USAAREUE Director of GED
APO New York 09403

1 DR. RALPH CANTER
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 DR. RALPH DUSEK
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Ed Johnson
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Milton S. Katz
Individual Training & Skill Evaluation Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
ATTN: PERI-OK
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Robert Ross
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Director, Training Development
U.S. Army Administration Center
ATTN: Dr. Sherrill
Ft. Benjamin Harrison, IN 46218

1 Dr. Frederick Steinheiser
U. S. Army Reserch Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/PED
Brooks AFB, TX 78235

1 Air University Library
AUL/LSE 76/443
Maxwell AFB, AL 36112

1 Dr. Philip De Leo
AFHRL/TT
Lowry AFB, CO 80230

1 DR. G. A. ECKSTRAND
AFHRL/AS
WRIGHT-PATTERSON AFB, OH 45433

1 CDR. MERCER
CNET LIAISON OFFICER
AFHRL/FLYING TRAINING DIV.
WILLIAMS AFB, AZ 85224

1 Dr. Ross L. Morgan (AFHRL/ASR)
Wright -Patterson AFB
Ohio 45433

1 Dr. Roger Pennell
AFHRL/TT
Lowry AFB, CO 80230

1 Personnel Analysis Division
HQ USAF/DPXXA
Washington, DC 20330

1 Research Branch
AFMPC/DPMYP
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/PED
Brooks AFB, TX 78235

1 Dr. Marty Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Jack A. Thorpe, Capt, USAF
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Brian K. Waters, LCOL, USAF
Air University
Maxwell AFB
Montgomery, AL 36112

Marines

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 MCDEC
Quantico Marine Corps Base
Quantico, VA 22134

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 MR. JOSEPH J. COWAN, CHIEF
PSYCHOLOGICAL RESEARCH (G-P-1/62)
U.S. COAST GUARD HQ
WASHINGTON, DC 20590

1 Dr. Thomas Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, VA 22314
Attn: TC

1 Dr. Dexter Fletcher
ADVANCED RESEARCH PROJECTS AGENCY
1400 WILSON BLVD.
ARLINGTON, VA 22209

1 Military Assistant for Training and Personnel Technology
Office of the Under Secretary of Defense for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 MAJOR Wayne Sellman, USAF
Office of the Assistant Secretary of Defense (MRA&L)
3B930 The Pentagon
Washington, DC 20301

Civil Govt

1 Dr. Susan Chipman
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. William Gorham, Director
Personnel R&D Center
U.S. Civil Service Commission
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph I. Lipson
Division of Science Education
Room W-638
National Science Foundation
Washington, DC 20550

1 Dr. John Mays
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Arthur Melmed
National Intitute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Andrew R. Molnar
Science Education Dev. and Research
National Science Foundation
Washington, DC 20550

1 Dr. Lalitha P. Sanathanan
Environmental Impact Studies Division
Argonne National Laboratory
9700 S. Cass Avenue
Argonne, IL 60439

1 Dr. Jeffrey Schiller
National Institute of Education
1200 19th St. NW
Washington, DC 20208

1 Dr. Thomas G. Sticht
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Vern W. Urry
Personnel R&D Center
U.S. Civil Service Commission
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Erling B. Anderson
University of Copenhagen
Studiestraedt
Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Alan Baddeley
Medical Research Council
Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF
ENGLAND

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Warner Birice
Streitkraefteamt
Rosenberg 5300
Bonn, West Germany D-5300

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
600 Jay Street
Sacramento, CA 95819

1 Dr. David G. Bowers
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106

1 Dr. Robert Brennan
American College Testing Progr
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William Coffman
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

1 Dr. Allan M. Collins
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, Ma 02138

1 Dr. Meredith Crawford
Department of Engineering Administration
George Washington University
Suite 805
2101 L Street N. W.
Washington, DC 20037

1 Dr. Hans Cronbag
Education Research Center
University of Leyden
Boerhaavelaan 2
Leyden
The NETHERLANDS

1 MAJOR I. N. EVONIC
CANADIAN FORCES PERS. APPLIED RESEARCH
1107 AVENUE ROAD
TORONTO, ONTARIO, CANADA

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Gerhardt Fischer
Liebigasse 5
Vienna 1010
Austria

1 Dr. Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ross Greene
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

1 Dr. Alan Gross
Center for Advanced Study in Education
City University of New York
New York, NY 10036

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
Department of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Carl J. Jensema
Gallaudet College
Kendall Green
Washington, DC 20002

1 Dr. Arnold F. Kanarick
Honeywell, Inc.
2600 Ridgeway Pkwy
Minneapolis, MN 55413

1 Dr. John A. Keats
University of Newcastle
Newcastle, New South Wales
AUSTRALIA

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 LCOL. C.R.J. LAFLEUR
PERSONNEL APPLIED RESEARCH
NATIONAL DEFENSE HQS
101 COLONEL BY DRIVE
OTTAWA, CANADA K1A OK2

1 Dr. Michael Levine
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Cortona Drive
Santa Barbara Research Pk.
Goleta, CA 93017

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77025

1 Dr. Sam Mayo
Loyola University of Chicago
Chicago, IL 60601

1 Dr. Allen Munro
Univ. of So. California
Behavioral Technology Labs
3717 South Hope Street
Los Angeles, CA 90007

1 Dr. Melvin R. Novick
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analysis
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. STEVEN M. PINE
4950 Douglas Avenue
Golden Valley, MN 55416

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

1 MIN. RET. M. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 161
53 BONN 1, GERMANY

1 Dr. Peter B. Read
Social Science Research Council
605 Third Avenue
New York, NY 10016

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
12 Hill Hall
Columbia, MO 65201

1 Dr. Fred Reif
SESAME
c/o Physics Department
University of California
Berkely, CA 94720

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Donald Rubin
Educational Testing Service
Princeton, NJ 08450

1 Dr. Larry Rudner
Gallaudet College
Kendall Green
Washington, DC 20002

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

1 Dr. Kazao Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 982
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
Florida Technological University
Orlando, FL 32316

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 DR. ALBERT STEVENS
BOLT BERANEK & NEWMAN, INC.
50 MOULTON STREET
CAMBRIDGE, MA 02138

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Bureau of Social SCience Research
1990 M Street, N. W.
Washington, DC 20036

1 DR. THOMAS WALLSTEN
PSYCHOMETRIC LABORATORY
DAVIE HALL 013A
UNIVERSITY OF NORTH CAROL
CHAPEL HILL, NC 27514

1 Dr. John Wannous
Department of Management
Michigan University
East Lansing, MI 48824

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Dr. Wolfgang Wildgrube
Streitkraefteamt
Rosenberg 5300
Bonn, West Germany D-5300

1 Dr. Robert Woud
School Examination Department
University of London
66-72 Gower Street
London WCIE 6EE
ENGLAND

1 Dr. Karl Zinn
Center for research on Learning and Teaching
University of Michigan
Ann Arbor, MI 48104

Previous Publications

Proceedings of the 1977 Computerized Adaptive Testing Conference. July 1978.

Research Reports

Final Report: Bias-Free Computerized Testing. March 1979.
79-2. Effects of Computerized Adaptive Testing on Black and White Students. March 1979.
79-1. Computer Programs for Scoring Test Data with Item Characteristic Curve Models. February 1979.
78-5. An Item Bias Investigation of a Standardized Aptitude Test. December 1978.
78-4. A Construct Validation of Adaptive Achievement Testing. November 1978.
78-3. A Comparison of Levels and Dimensions of Performance in Black and White Groups on Tests of Vocabulary, Mathematics, and Spatial Ability. October 1978.
78-2. The Effects of Knowledge of Results and Test Difficulty on Ability Test Performance and Psychological Reactions to Testing. September 1978.
78-1. A Comparison of the Fairness of Adaptive and Conventional Testing Strategies. August 1978.
77-7. An Information Comparison of Conventional and Adaptive Tests in the Measurement of Classroom Achievement. October 1977.
77-6. An Adaptive Testing Strategy for Achievement Test Batteries. October 1977.
77-5. Calibration of an Item Pool for the Adaptive Measurement of Achievement. September 1977.
77-4. A Rapid Item-Search Procedure for Bayesian Adaptive Testing. May 1977.
77-3. Accuracy of Perceived Test-Item Difficulties. May 1977.
77-2. A Comparison of Information Functions of Multiple-Choice and Free-Response Vocabulary Items. April 1977.
77-1. Applications of Computerized Adaptive Testing. March 1977.
Final Report: Computerized Ability Testing, 1972-75. April 1976.
76-5. Effects of Item Characteristics on Test Fairness. December 1976.
76-4. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. June 1976.
76-3. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. June 1976.
76-2. Effects of Time Limits on Test-Taking Behavior. April 1976.
76-1. Some Properties of a Bayesian Adaptive Ability Testing Strategy. March 1976.
75-6. A Simulation Study of Stradaptive Ability Testing. December 1975.
75-5. Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975.
75-4. A Study of Computer-Administered Stradaptive Ability Testing. October 1975.
75-3. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975.
75-2. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975.
75-1. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975.
74-5. Strategies of Adaptive Ability Measurement. December 1974.
74-4. Simulation Studies of Two-Stage Ability Testing. October 1974.
74-3. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974.
74-2. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974.
74-1. A Computer Software System for Adaptive Ability Measurement. January 1974.
73-3. The Stratified Adaptive Computerized Ability Test. September 1973.
73-2. Comparison of Four Empirical Item Scoring Procedures. August 1973.
73-1. Ability Measurement: Conventional or Adaptive? February 1973.

Copies of these reports are available, while supplies last, from:
Psychometric Methods Program, Department of Psychology
N660 Elliott Hall, University of Minnesota
75 East River Road, Minneapolis, Minnesota 55455